CHAPTER 6

STATISTICAL CONCEPTS

INTRODUCTION

As we mentioned in Chapter 5, our assumptions about a given testing situation lead us to the choice of a mathematical model to characterize the reliability of a system. However, we cannot determine the actual reliability of the system using the model until the parameters of the model, p for the binomial model and A (or θ) for the Poisson or exponential model, have been specified. The values of the parameters are never known with absolute certainty. As a consequence, some form of sampling or testing is required to obtain estimates for these parameters. The quality of the estimates is, of course, directly related to the quality and size of the sample.

POINT ESTIMATES

point estimates represent a single "best guess" about model parameters, based on the sample data. A distinguishing symbol commonly is used to designate the estimate of a parameter. Most commonly, a caret or "hat" is used to designate point estimates (e.g., $\hat{\theta}$, $\hat{R}(x)$, $\hat{\lambda}$). Quite often, and for our purposes, the caret further indicates that the estimator is a maximum likelihood estimator; that is, it is the most likely value of the parameter of the model which is presumed to have generated the actual data.

There are criteria other than maximum likelihood used for a single "best guess." One other is unbiasedness. For an estimator to be unbiased, we mean that, in the long run, it will have no tendency toward estimating either too high or too low. The point estimates which we propose for p in the binomial model and for A in the Poisson and exponential models are both maximum likelihood and unbiased.

CONFIDENCE STATEMENTS

Point estimates represent a single "best guess" about parameters, based on a single sample. The actual computed values could greatly overestimate or underestimate the true reliability parameters, particularly if they are based on a small amount of data. As an example, suppose that 20 rounds of ammunition were tested and 18 fired successfully.

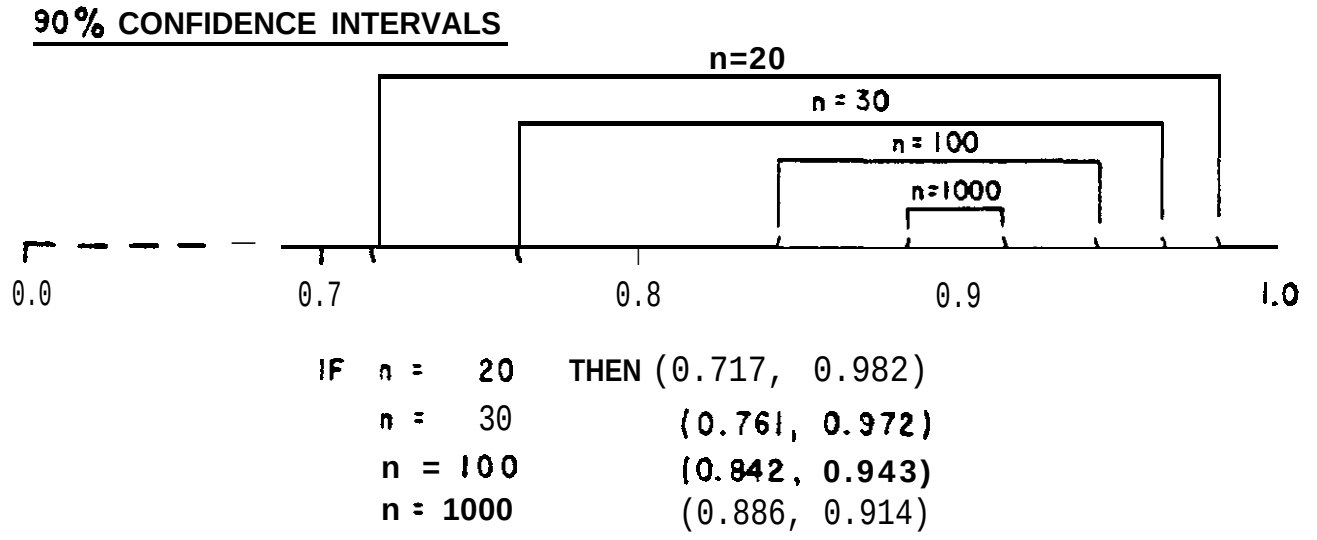
The maximum likelihood and unbiased estimate of reliability is $\hat{R} = 18/20 = 0.9$. In other words, the system most likely to have generated 18 successes is one whose reliability is 0.9. Note that 0.9 is the percentage of successes actually observed in the sample. However, a system whose true reliability is somewhat less than or somewhat more than 0.9 could reasonably have generated this particular data set.

We use confidence limits to address how high or low the value of a parameter could reasonably be. A 90% confidence interval for reliability is: $0.717 < R < 0.982$. In other words, if being reasonable signifies being 90% confident of

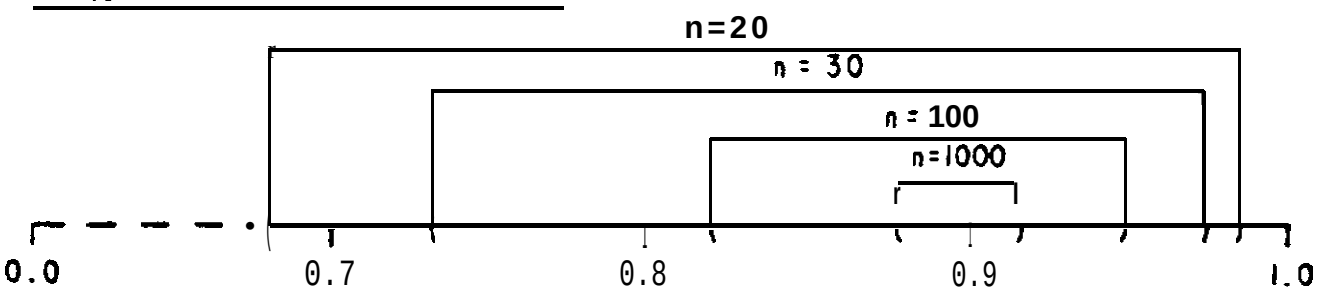
being right, then it is unreasonable to consider that a system whose reliability is actually less than 0.717 or one whose reliability is actually more than 0.982 generated the 18 successful rounds . When we desire to be more confident, say 95% confident, that our interval contains the true system reliability, we widen our interval, i.e. , we expand the group of systems considered to have reasonably generated the data. A 95% confidence interval for the reliability of our example system is: $0.683 < R < 0.988$. Since we are now allowing for the possibility that the system reliability could be a little lower than 0.717 -- namely, as low as 0.683 -- or a little higher than 0.982 -- namely, as high as 0.988 -- we can now afford to be more confident that our interval indeed contains the *true* value. For a fixed amount of testing, we can only increase our confidence by widening the interval of reasonable values.

Suppose that we desire to reduce the size of the interval while maintaining the same level of confidence or to increase the level of confidence while maintaining approximately the same size interval. Either of these objectives is accomplished through increased testing, i.e., taking a larger sample. If the system test had resulted in 27 successful firings out of 30 attempts (vice 18 out of 20) , the point estimate is still 0.9. However, the 90% confidence interval for system reliability is: $0.761 < R < 0.972$. The length of this interval represents a 20% reduction in the length of the 90% confidence interval resulting from our test of 20 units. The 95% confidence interval for system reliability is: $0.734 < R < 0.979$. This interval represents an 8% reduction in size, but our confidence has increased to 95%. Figure 6-1 graphically portrays the effect on interval length induced by changing confidence levels or increasing sample size.

FIGURE 6-1 CONFIDENCE INTERVALS



IF n = 20 THEN (0.717, 0.982)
n = 30 (0.761, 0.972)
n = 100 (0.842, 0.943)
n = 1000 (0.886, 0.914)

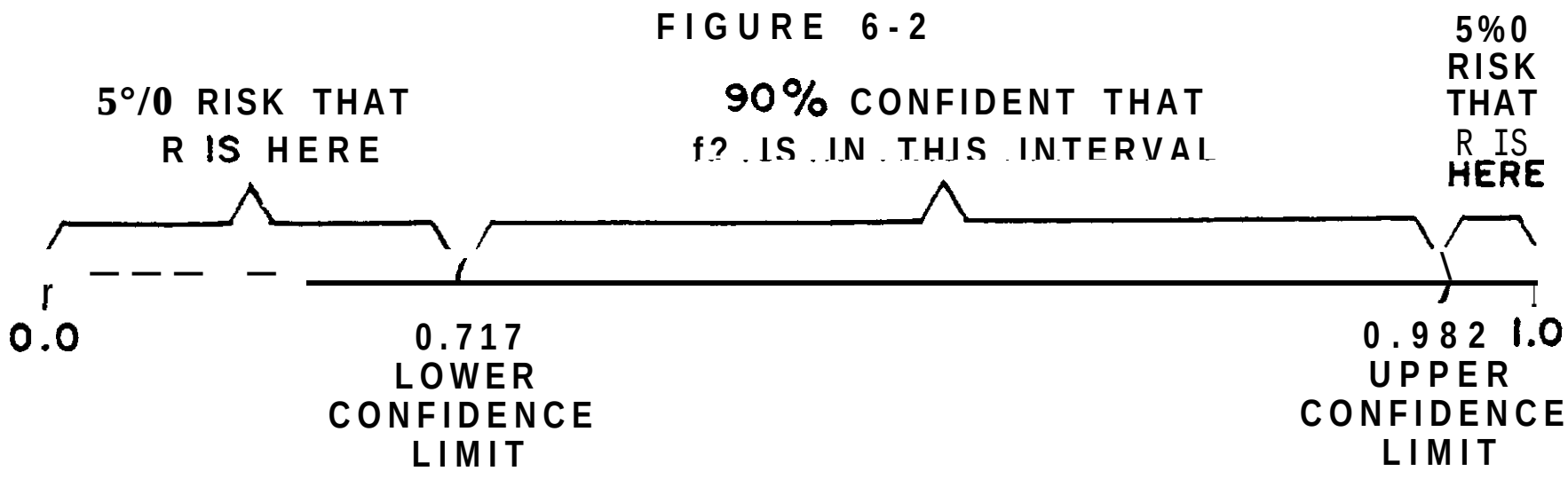


IF n = 20 THEN (0. 683, 0.988)
n = 30 (0. 734, 0.979)
n = 100 (0. .922, 0.951)
n = 1000 (0.882, 0.918)

A cautious, conservative person who buys safe investments, wears a belt and suspenders, and qualifies his statements carefully is operating on a high-confidence level. He is certain he won't be wrong very often. If he is wrong once in 100 times, he is operating on a 99% confidence level. A less conservative person who takes more chances will be wrong more often, and hence he operates o-n a lower confidence level. If he is wrong once in 20 times, he is operating on a 95% confidence level. The confidence level, therefore, merely specifies the percentage of the statements that a person expects to be correct. If the experimenter selects a confidence level that is too high, the test program will be prohibitively expensive before any very precise conclusions are reached. If the confidence level is too low, precise conclusions will be reached easily, but these conclusions will be wrong too frequently, and, in turn, too expensive if a large quantity of the item is made on the basis of erroneous conclusions. There is no ready answer to this dilemma.

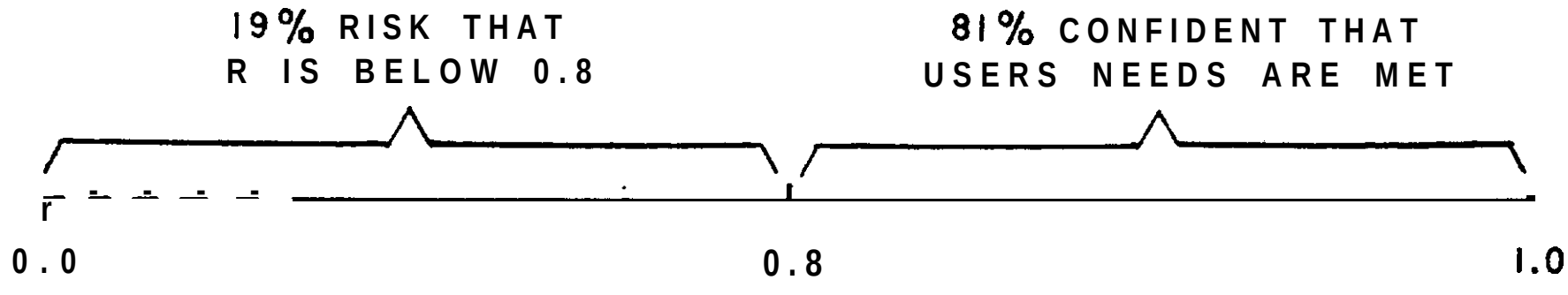
We can interpret confidence statements using the concept of risk. With a 90% confidence statement, there is a 10% risk; with a 99% confidence statement, there is a 1% risk. Confidence intervals generally are constructed so that half of the total risk is associated with each limit or extreme of the interval. Using this approach with a 90% interval for reliability, there is a 5% risk that the true reliability is below the lower limit and also a 5% risk that the true reliability is above the upper limit. We can therefore state for the example system with 18 of 20 successes that we are 95% confident that: $R > 0.717$. This is a lower confidence limit statement. We are also 95% confident that: $R < 0.982$. This is an upper confidence limit statement. See Figure 6-2.

FIGURE 6-2



The classical textbook approach to confidence intervals has been to specify the desired confidence level and determine the limit associated with this confidence level. This approach creates a twofold problem. First, the desired confidence level has to be determined. Second, the limits that are generated are generally not, in themselves, values of direct interest. A very practical modification is to determine the level of confidence associated with a predetermined limit value. For example, the minimum value of a reliability measure that is acceptable to the user is a logical lower limit. The confidence in this value can then be interpreted as the assurance that the user's needs are met. See Figure 6-3.

FIGURE 6-3 CONFIDENCE INTERVALS- ACCEPTABLE LOWER LIMITS



The confidence level for a lower limit of 0.8 is 81%. A system reliability of 0.8 is the user' s minimum acceptable value (MAV) .

HYPOTHESIS TESTING

While confidence limits are generally used to define the uncertainty of a parameter value, an alternative approach is hypothesis testing. Both approaches essentially give the same information. Hypothesis testing can be used to distinguish between two values or two sets of values for the proportion of failures in a binomial experiment, or for the failure rate in a Poisson/ exponential experiment. Let us examine hypothesis testing using a binomial example. Typically, for a binomial experiment, it is hypothesized that the probability of failure, p, is a specified value. While there is seldom any belief that p is actually equal to that value, there are values of p which would be considered unacceptable in a development program. These unacceptable values are specified in an alternative hypothesis. Consider the following examples.

(1) One-Sided Tests

H_0: p = 0.3 (Null Hypothesis)

H_1: p > 0.3 (Alternative Hypothesis)

In Case (1) , the evaluator hopes that p is no more than 0.3. He considers a p of more than 0.3 to be unacceptable. This is a classical one-sided test. Another type of one-sided test has the alternative hypothesis p < 0.3.

(2) Two-Sided Tests

H_0: p = 0.3

H_1: p # 0.3

In Case (2) , the evaluator hopes that p is approximately 0.3. Values of p much larger than or much smaller than O. 3 are unacceptable. This is a classical two-sided test.

(3) Simple vs. Simple Tests

H_0: p = 0.3

H_1: p = 0.5

In Case 3, the evaluator hopes that p is no more than 0.3. He considers a p of more than 0.5 to be unacceptable. The region between 0.3 and 0.5 is an indifference region in that it represents acceptable but not hoped for values. This is actually a classical simple versus simple test. This type of test is treated extensively and exclusively in Chapter 8.

In order to conduct a statistical test of hypothesis, the following steps are employed:

1. The hypothesis, null and alternative, are specified. For our purposes, the null hypothesis is the contractually specified value (SV) and the alternative hypothesis is the minimum acceptable value (MAV).

2. A sample size, n, is determined. This value must be large enough to allow us to distinguish between the SV and MAV. Chapter 8 is devoted to procedures for determining a sufficiently large value of n.

3. An accept/reject criterion is established. For our purposes, this criterion is established by specifying a value c, which is the maximum number of failures permitted before a system will be rejected.

4. The sample is taken and the hypothesis is chosen based upon the accept/reject criterion. If c or fewer failures occur, we accept the system. If more than c failures occur, we reject the system.

PRODUCER ' S AND CONSUMER'S RISKS

There are two possible errors in making a hypothesis-testing decision. We can choose the alternative hypothesis, thereby rejecting the null hypothesis, when, in fact, the null hypothesis is true. The chance or probability of this occurring is called the producer's risk, α. On the other hand, we can choose the null hypothesis, i.e., accept it as reasonable, when in fact the alternative hypothesis is true. The chance or probability of this occurring is termed the consumer's risk, β. See Chapter 8 for an additional discussion of this topic.

Consider the following: A system is under development. It is desired that it have a 300-hour MTBF. However, an MTBF of less than 150 hours is unacceptable, i.e. , the MAV is 150 hours. How would we set up a hypothesis test to determine the acceptability of this new system? Our null hypothesis (desired value) is that the MTBF is 300 hours. Our alternative hypothesis (values of interest) is that the MTBF has a value which is less than 150 hours. To decide which hypothesis we will choose, we determine a test exPosure and decision criterion. The α risk (producer's risk) is the probability that the decision criterion will lead to a rejection decision when in fact the system meets the specification of 300 hours MTBF. The β risk (consumer' s risk) is the probability that the decision criterion will lead to an acceptance decision when in fact the system falls short of the 150 hours MTBF.

For a given test, the decision criteria can be altered to change the α and β risks. Unfortunately, a decision criterion which decreases one automatically increases the other. The only way to decrease both risks is to increase the test exposure, that is, the number of test hours. We address this area below in Chapter 8, "Reliability Test Planning."

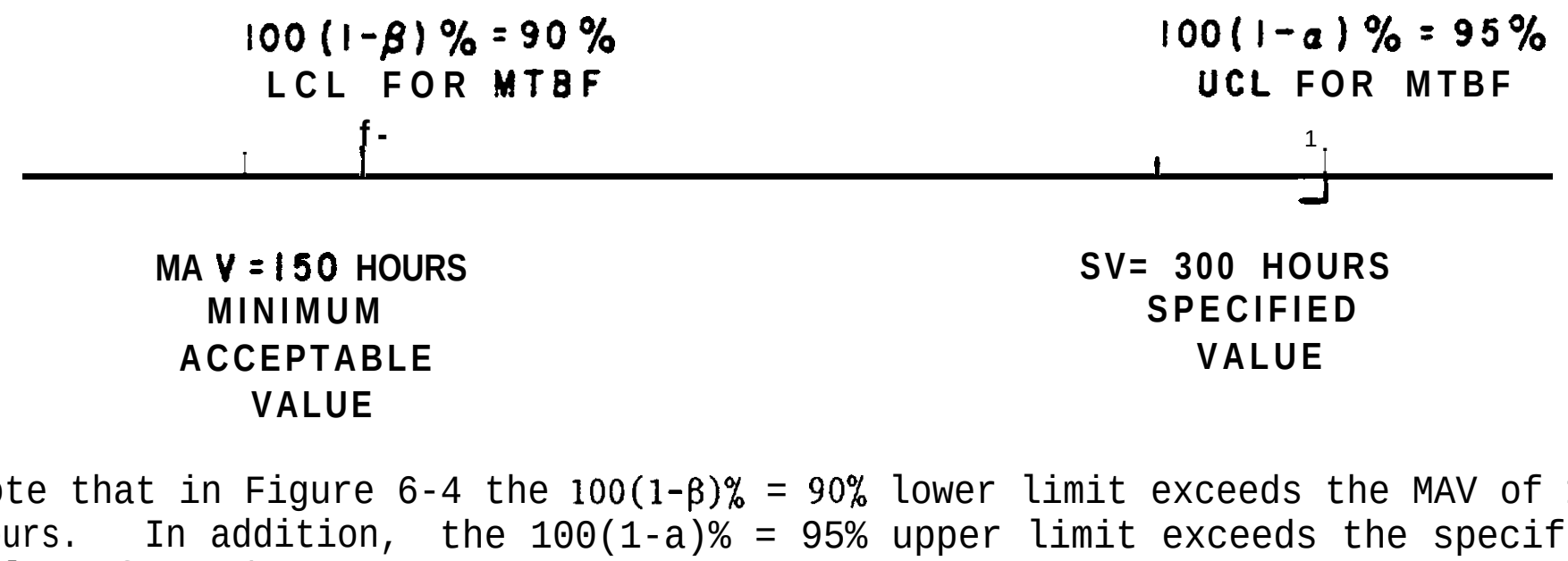
INTERFACE BETWEEN HYPOTHESIS TESTING AND CONFIDENCE STATEMENTS

In both test planning and data analysis situations, either hypothesis testing or confidence statements provide an avenue of approach. The interface between the two approaches can be best understood through the following example.

Suppose α is the desired producer's risk ($\alpha = 0.05$) for the specified MTBF of 300 hours. Suppose further that β is the desired consumer's risk ($\beta = 0.1$) for the minimum acceptable MTBF of 150 hours. The hypothesis testing approach determines a required sample size and a specified accept/reject criterion. We show how the same information can be obtained through confidence statements in the following two cases. The abbreviations LCL and UCL represent Lower Confidence Limit and Upper Confidence Limit, respectively.

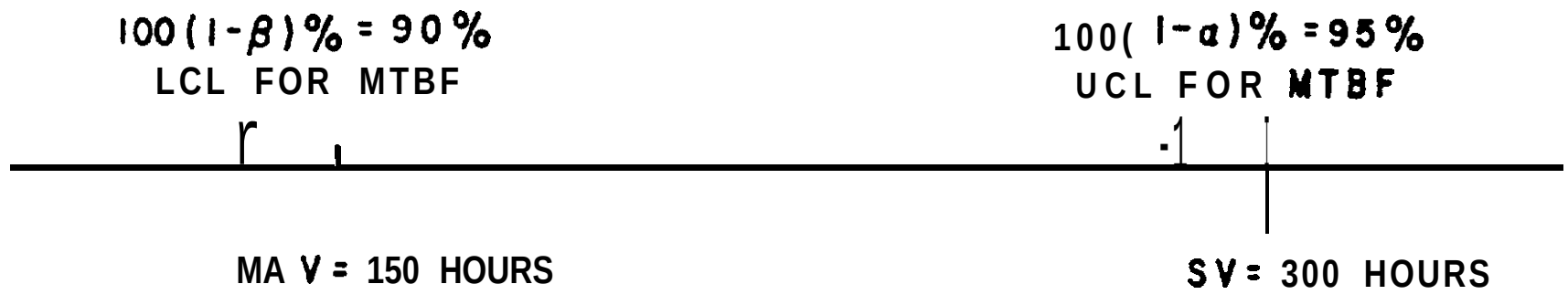
Note that the distance between the upper and lower limits is the same as the distance between the SV and the MAV. When this is the case we shall always be able to make a clear-cut decision and the risks associated with the decision will be as specified at the outset of testing.

FIGURE 6-4 ACCEPTANCE DECISION



Note that in Figure 6-4 the 100(1-β)% = 90% lower limit exceeds the MAV of 150 hours. In addition, the 100(1-a)% = 95% upper limit exceeds the specified value of 300 hours. The consumer is 90% confident that the 150-hour MAV has been met or exceeded and the producer has demonstrated that the system could reasonably have a 300-hour MTBF. Consequently, we would make the decision to accept the system.

FIGURE 6-5 REJECTION DECISION



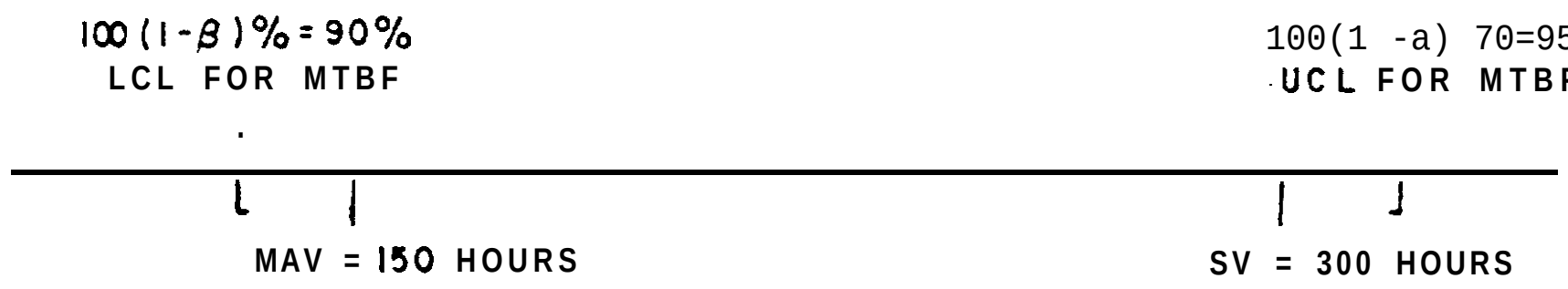
Note that in Figure 6-5 the $100(1-\beta)\% = 90\%$ lower limit falls below the MAV of 150 hours. In addition, the 100(1-cY)% = 95% upper limit falls below the SV of 300 hours. Therefore, the true MTBF could reasonably be below 150 hours and the producer has not demonstrated that an MTBF of 300 hours is reasonable. Consequently, we make the decision to reject the system.

TEST EXPOSURE

Perhaps one of the most important subjects to be considered in the evaluation of RAM characteristics is the subject of test exposure. The term "test exposure" refers to the amount (quantity and quality) of testing performed on a system or systems in an effort to evaluate performance factors. In Chapter 10, we discuss the qualitative aspects of test exposure which should be considered by the test designer. The primary purpose of Chapter 8, "Reliability Test Planning," is to document procedures which ensure that the quantitative aspects of test planning are adequate.

Recall the comment we made in the previous section to the effect that the difference in the distance between the upper and lower confidence <limits was equal to the difference in the distance between the SV and the NAV. When this condition is achieved, we have obtained the most efficient test exposure for the stated requirements and risks. Examples of situations where test exposure is inadequate or excessive are given below. See Case Study 6-2 for an illustration of the evaluation of a proposed test exposure.

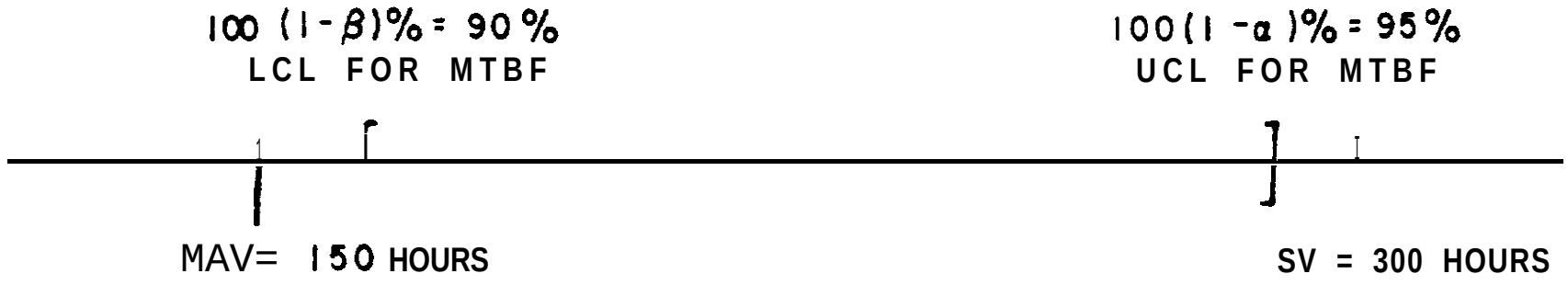
FIGURE 6-6 INADEQUATE TEST DURATION



Note that in Figure 6-6 the $100(1-\beta)\% = 90\%$ lower limit falls below the MAV of 150 hours. The 100(1-cY)% = 95% upper limit exceeds the SV of 300 hours. The true MTBF could reasonably be below 150 hours or above 300 hours. Test exposure is insufficient to discriminate between the MAV of 150 hours and the SV of 300 hours with the required risk levels of 10% and 5%. If we reject the system, the producer can legitimately claim that an MTBF of 300 hours is reasonable for his system. On the other hand, if we accept the system, we may be fielding an inadequate system.

Note that in Figure 6-7 the 100(1-β)% = 90% lower limit exceeds the MAV of 150 hours. The 100(1-cY)% = 95% upper limit falls below the SV of 300 hours. The consumer has 90% confidence that the 150-hour MAV has been met or exceeded. However, the producer has not demonstrated the specified 300-hour MTBF. The test exposure is more than required to obtain the risks of 10% and 5% for the stated values of MAV and SV. Since the MAV has been met or exceeded, we will probably accept the system. We may have paid a premium to obtain information that allowed us to construct a confidence interval more narrow than required.

FIGURE 6-7 EXCESSIVE TEST DURATION



CASE STUDY NO. 6-1

Background

A contract for a new electronic system specifies an MTBF of 1000 hours. The minimum acceptable value is 500 hours MTBF. A design qualification test is to be conducted prior to production. The test risks are to be 20% for consumer and 10% for producer.

Determine

Describe the events which lead to acceptance or rejection of the system-

Solution

In accordance with procedures defined in Chapter 7, "Reliability Data Analysis," the appropriate hypothesis test is set up, the sample is taken, and the data are analyzed.

The Positive Chain of Events

1. The contractor has met (or exceeded) an MTBF of 1000 hours.

2. There is (at least) a 0.90 probability of "passing" the test.

3. "Passing" the test will give the user (at least) 80% confidence that the MAV of 500 hours MTBF has been exceeded.

4. The user is assured that his needs have been met.

The Negative Chain of Events

1. The contractor has met an MTBF of 500 hours (or less).

2. There is (at least) a 0.80 probability of "failing" the test.

3. "**Failing**" the test gives the procuring activity (at least) 90% confidence that the contractually obligated SV of 1000 hours MTBF has not been met.

4. The procuring activity is assured that the contractual obligations have not been met.

CASE STUDY NO. 6-2

Background

The specified MTBF of a targeting system is 500 hours and the minimum acceptable MTBF is 400 hours. The contractor has proposed a development test consisting of 6000 hours on the initial prototype system and 2000 hours on a second prototype system which will contain some minor engineering advances.

The proposed test plan of 8000 hours can distinguish between the SV of 600 hours and the MAV of 400 hours for consumer' s and producer' s risks of slightly over 20%. If the producer is willing to accept a 30% producer's risk, the proposed plan will yield a 12% consumer' s risk.

Determine

Comment on the adequacy of the proposed test.

Solution

These risks seem to be larger than should be considered for an important system. The test exposure seems to be inadequate for the following reasons:

- Test time is not of sufficient length.
- Prototypes are not identical . Test time on the second prototype may not be long enough to determine if the design improvements increase reliability.
- Only two systems on test may be insufficient. Ideally, more systems should be used for shorter periods of time.

A test plan having four systems accumulating about 4000 hours each will yield producer and consumer risks of just over 10%. A further benefit is that using four systems and operating them for a period of time about 10 times the minimum MTBF should paint a pretty clear picture of the system capability throughout a significant part of its expected age.

Note : Chapter 8 will present the analytical tools required to evaluate the above test plan. Our objective here is to qualitatively review the various aspects of a statistically relevant test program.